Corpógrafo

Um ambiente livre para o ensino e desenvolvimento de Terminologia

Luís Sarmento las@fe.up.pt

Auto-Apresentação

- Aluno de doutoramento Engenharia Informática
  - Faculdade de Engenharia da Universidade do Porto
  - Análise Semântica Robusta do Português
- Orientadores:
  - Prof. Eugénio Oliveira (LIACC: www.fe.up.pt/~eol)
  - Dra. Diana Santos (Linguateca: www.linguateca.pt)
- Últimos 4 anos tenho colaborado com a Linguateca
  - Pólo do Porto (Prof. Belinda Maia)
  - Projectos: Corpógrafo, REPENTINO, SIEMÊS…
- Mais informação sobre mim: www.fe.up.pt/~las/

Linguateca, um projecto para o português

- Centro de recursos distribuído para o processamento computacional da língua portuguesa
- Projecto financiado pela FCT através do POSI (2000-2006)
- Primeiro pólo no SINTEF ICT, Oslo, começou em 2000 (actividade no SINTEF começou em 1998 com o projecto **Processamento Computacional do Português**)

Modelo IRA

- **I**nformação
- **R**ecursos
- **A**valiação

Linguateca num relance, www.linguateca.pt

- > 1000 links Mais de 1.500.000 visitas ao site
- AC/DC, CETEMPúblico, COMPARA … Recursos valiosos para o processamento do português
- *Morfolimpíadas, parte portuguesa do CLEF, HAREM* Avaliação conjunta para o português

- Recursos públicos
- Incentivar a investigação e colaboração
- Medida e comparação formal

- Uma língua, muitas culturas
- Cooperação usando a Web
- Não à adaptação directa das aplicações para o inglês

Contacto: Diana Santos@sintef.no

LIACC -- NIAD&R

- LIACC unidade de investigação criada em 1998 na Universidade Porto
- Composta por 3 Núcleos:
  - NCC- Ciências da Computação - Fac. de Ciências
  - NIAAD- Int, Artificial e Análise de Dados - Fac. Economia
  - **NIAD&R- Robótica, Int. Artificial Distribuida -Fac. Engenharia**
- Comité Organizador :
  - NCC: Prof. Miguel Filgueiras, Dr. Luís Damas
  - NIAAD: Prof. Pavel Brazdil
  - NIAD&R: Prof. Eugénio Oliveira

NIAD&R

- Tópicos de Investigação:
  - Sistemas Multi-Agente
  - Tecnologias de Negócio Electrónico
  - Aprendizagem Automática
  - Robótica
  - Ontologias
  - *Análise Semântica do Português*
- Contacto:
  - Prof Eugénio de Oliveira eco@fe.up.pt

Esta apresentação…

- Dividida em duas partes:
- Parte 1: Mais "teórica"
  - Enquadramento Geral
  - Corpógrafo, o que é? Como surgiu? Como funciona? Etc..
- Parte 2: Prática
  - Mexer um pouco na ferramenta on-line

Introdução (1)

- Algumas questões que vou tentar responder:
  - O que é o Corpógrafo?
  - Porque é que eu estou a falar sobre o Corpógrafo?
  - Para quem pode ser interessante esta apresentação?
  - Como surgiu o Corpógrafo?
  - Como funciona o Corpógrafo?
  - Quem são os utilizadores do Corpógrafo?
  - Quais são os pontos fortes e fracos do Corpógrafo?

Introdução (2)

- O Corpógrafo tem sido desenvolvido por uma equipa grande e muito distinta:
  - Ideia Original: Belinda Maia
  - Orientação Conceptual: Diana Santos
  - Arquitectura de Software: Luís Sarmento
  - Programação: Luís Sarmento (ex) & Luís Cabral
  - Recursos Ling. e Documentação: Ana Pinto; Débora Oliveira
  - Portabilidade e Qualidade: Rui Vilela & António Silva
- Formações diferentes:
  - Linguístas, Engenheiros, Tradutores, Cientistas de Comp.

Introdução (3)

- Esta apresentação traduz uma visão pessoal
  - de um Engenheiro
  - orientação para uma visão técnica sobre o assunto
- Não será uma palestra técnica
  - Mas tentará mostrar o que se passa por dentro do Corpógrafo e explicar os desafios da suas construção
- Se a apresentação fosse feita por outro dos participantes na construção do Corpógrafo:
  - A visão seria necessariamente diferente
  - A diferença de visões foi enriquecedora para o Corpógrafo

O que é o Corpógrafo?

- Dividir a questão em 3 outras questões:
  - O que é que se queria que o Corpógrafo fosse?
    - Necessidades iniciais
  - O que é o Corpógrafo se tornou efectivamente?
    - O que foi possível construir e como os utilizadores influenciaram as nossas ideias
  - O que é que o Corpógrafo pode vir a ser?
    - Como pode ser melhorado e o que necessita para que isso aconteça

O que é que se queria que o Corpógrafo fosse?

- Tentativa de suprimir lacunas no acesso a Corpora
  - É impossível que as entidades oficiais forneçam corpora para todo o tipo de utilizadores
  - Mas os utilizadores poderiam eles próprios fazer os seus próprios corpora:
    - Conceito "Do it yourself corpora" (Belinda Maia, 1997)
- Corpógrafo:
  - sistema de construção e pesquisa de Corpora
    - Monolingue, essencialmente técnico
    - COMPARÁVEL bilingue
- Auxílio para estudo de tradução (técnica!)

O que é o Corpógrafo se tornou efectivamente?

- Ferramenta que permite:
  - a construção de corpora monolingue
    - A partir de vários tipos de ficheiros
  - pesquisa de concordâncias, em janela, etc…
  - pesquisa semi-automática de terminologia
    - Armazenamento em BD (exportável em XML)
    - Posterior pesquisa semi-automática de definições / relações

O que é o Corpógrafo se tornou efectivamente? (2)

- Essencialmente utilizada:
  - no **ensino de terminologia** por várias Universidades
  - em alguns (ainda poucos) **projectos de terminologia** com parceiros dentro da Universidade do Porto
  - outros usos que deconhecemos
    - Há mais de 700 utilizadores inscritos (embora mais de metade sejam utilizadores que só visitam um Corpografo por um período de tempo reduzido)
- Permitiu criar um pequena comunidade de utilizadores interessados em corpora…
  - Já voltaremos a isto…

O que é que o Corpógrafo pode vir a ser?

- Depende de muitos factores:
  - Recursos humanos disponíveis
  - Interesse da comunidade
  - Capacidade real de execução
  - …
- Mas, acima de tudo, o Corpógrafo pode ser um elemento agregador da comunidade que trabalha com:
  - Terminologia (centrada no Português)
  - Tradução (centrada no Português)
- Voltaremos também a este assunto…

Porque é que ***eu*** estou a falar sobre o Corpógrafo?

- Participei no desenvolvimento do Corpógrafo:
  - Mas já não estou ligado ao seu desenvolvimento
- A experiência de desenvolvimento foi extremamente interessante:
  - Fusão de perspectivas: difícil mas útil
- Interessante poder partilhar com outros grupos a experiência de desenvolvimento
- Conhecer alguns utilizadores do Corpógrafo
  - Ouvir queixas! Ouvir sugestões para passar aos actuais responsáveis

Porque é que ***eu*** estou a falar sobre o Corpógrafo? (2)

- Mas não venho "vender" o Corpógrafo! :)
- Corpógrafo tem excesso de utilizadores:
  - Cresceu mais rápido do que nós conseguimos "acompanhar"
- Actualmente, a equipa de desenvolvimento está:
  - Replicar servidores (dividir os utilizadores)
  - Tentar tornar o Corpógrafo mais facilmente instalável por outras instituições
  - Melhorar a documentação de instalação e de utilizador
- Tudo isto demora o seu tempo…
  - por enquanto a capacidade do Corpógrafo está limitada…

Para quem pode ser interessante esta apresentação?

- Para Linguístas, Tradutores, Terminólogos
  - Corpógrafo **tem sido** uma ferramenta útil no **ensino** de Linguística de Corpora e Terminolgia
  - Corpógrafo **pode ser** uma ferramenta útil em **projectos** nestes domínios
- Para Engenheiros desenvolvendo software para LC:
  - Aproximação aos problemas de Linguística
  - Filosofia semi-automática das soluções
  - Análise de Requisitos motivada por necessidades práticas
    - não por aquilo que o "engenheiro pensa que os Linguístas querem"

Está tudo motivado?

- Restantes questões:
  - Como surgiu o Corpógrafo?
    - Alguma História
  - Como funciona o Corpógrafo?
    - Arquitecura
    - Fluxo de Funcionamento
  - Quem são os utilizadores do Corpógrafo?
    - Algumas estatisticas
  - Quais são as vantagens e os problemas do Corpógrafo?!

Como surgiu o Corpógrafo?

- O Corpógrafo foi surgindo…
  - motivado por necessidades dos utilizadores
  - a partir do início de 2003
- Primeiras GRANDES necessidades:
  - Como podem os utilizadores criar os seus próprios corpora?
  - Como podem usar o texto dos ficheiros PDF? Word?…
  - Como utilizar o texto em ferramentas como WordSmith
- Esta era a necessidade mais "básica":
  - Não interessa criar ferramentas avançadissimas se o "básico" não estiver criado

Era A.C. (Antes do Corpógrafo)

- Construimos o EXTEX: EXtractor de TEXto
  - talvez o ponto-chave do futuro Corpógrafo
- Interface Web para vários utilitários Unix de extração de texto a partir de ficheiros
  - PDF, PS, Word, RTF e HTML
- Um único ponto de acesso para o utilizador
  - Escondia todos os detalhes técnicos de funcionamento dos vários utilitários
- Texto podia depois ser utilizado em ferramentas como o WordSmith, etc:
  - Pesquisa de Concordâncias, etc…

EXTEX - Impacto

- Os utilizadores reagiram muito bem à ideia
  - Carregavam no servidor 1 ficheiro
  - Descarregavam texto
- Mas foi aí que começaram também a exigir mais:
  - "Eu não quero trabalhar só com um ficheiro!"
  - "Eu quero trabalhar com vários corpora!"
  - "Eu não tenho acesso ao WordSmith: é obrigatório comprar licenças? Onde posso arranjar alternativa?"
  - "Porque é que não posso fazer pesquisas directamente no EXTEX?"

Calma, calma, Sr. Utilizador!!

- Decidiu-se criar um projecto maior, "filho do Extex":
  - Foi baptizado "Gestor de Corpora"!
  - Madrinhas: Belinda Maia & Diana Santos
  - Padrinho: Luís Sarmento
- Primeira versão em meados de 2003
  - Apresentado originalmente no CL2003

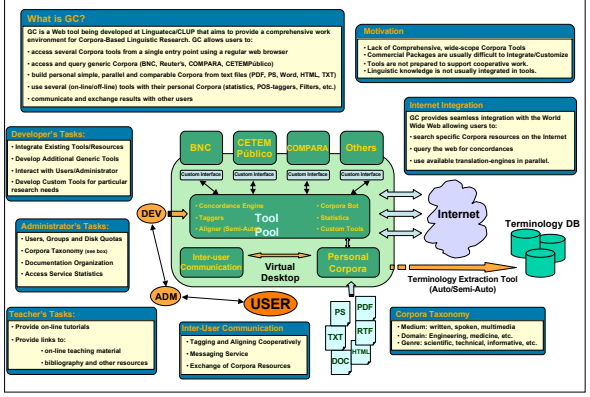
Gestor de Corpora

- Cada utilizador tinha uma conta num servidor:
  - Autenticação por palavra passe
  - As áreas eram pessoais
- O Utilizador podia carregar ficheiros em vários formatos para a sua conta
  - O texto era extraído dos PDF, PS, Word, RTF e HTML
  - O utilizador podia "limpar" o texto
- Os textos podiam ser agrupados para formar "Corpora Pesquisáveis"
  - O motor de indexação e pesquisa era o IMS-CWB
  - Permitia pesquisas de concordância e KWIC

Gestor de Corpora (2)

- GC era um interface web de fácil utilização para:
  - vários utilitários de extracção de texto livres
  - um motor de indexação e pesquisa de texto
  - uma área em disco onde guardar corpora
- GC tinha uma Interface de Administração
  - Criar utilizadores, alterar definições do sistema
- Problemas resolvidos:
  - todo o texto era para utilização pessoal, não há problemas de copyright dos texto
  - acesso Web: o utilizador não precisa de instalar nada
  - Fácil utilização nas aulas, em casa, sem custos…

GC no CL2003



Problemas resolvidos… problemas criados…

- Fácil para utilizador, difícil para engenheiro:
  - Difícil integração das ferramentas exteriores
    - Ainda um problema na actual versão
  - Dezenas de interfaces web para manter!!
- Utilizador:
  - "Se eu já tenho os meus corpora e se já posso pesquisar texto, então era mesmo mesmo bom poder pesquisar terminologia técnica!!!"
    - Lá vamos nós outra vez…

Pesquisa de terminologia

- O que era necessário:
  - Encontrar terminologia em corpora técnico
  - Texto em vários idiomas possíveis (PT, EN, FR, ES,…)
  - Armazenar os termos numa pequena BD
- Terminologia parecia também oportunidade para avançar num dos objectivos iniciais:
  - Criação de Corpora Comparável Bilingue: terminologia bilingue alinhada ajudaria a "alinhar" esse corpora
- Pareceu também uma boa oportunidade para rebaptizar o sistema: nasceu o "Corpógrafo"

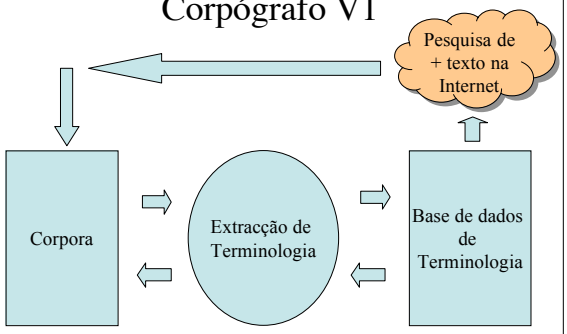
Filosofia Semi-Automática

- Ponto fundamental do Corpógrafo:
  - As ferramentas não precisam de fazer tudo automáticamente
  - As ferramentas precisam é de acelerar o trabalho do utilizador
- Do ponto de vista de engenharia facilita:
  - Os métodos não têm de ter um nível de 100%
  - Os algoritmos podem ser mais simples
- Objectivo:
  - **Ajudar** o utilizador na pesquisa e não **substituir** o utilizador

Extracção de Terminologia

- Sistema de extracção de terminologia que usa um método de "light parsing":
  - Funciona com base na pesquisa de n-gramas
  - Remove n-gramas que "parecem" incorrectos/mal-formados
  - Usa uma pequena lista de restrições léxicais para cada uma das linguas: PT, EN, FR, ES, IT
    - em vez de "parsers" (que não estariam disponíveis)
- Não sendo perfeito, serve para a tarefa:
  - Acelera o processo manual dezenas de vezes
  - E não implicou um acréscimo de dependências exteriores que tornariam um Corpógrafo mais complexo de manter…

Extracção de Terminologia no Corpógrafo V1



Corpógrafo V1

- Capacidades:
  - Preparação e Compilação de Corpora
  - Pesquisa de Concordâncias e KWIC
  - Extracção de Terminologia
    - Compilação de Bases de Dados
- Utilizadores eram essencialmente alunos do MTT da FLUP (Prof. Belinda Maia)

Problemas resolvidos… problemas criados… Parte 2

- "Mas se eu já compilo bases de dados de terminologia, então porque é que não pesquiso definições e relações entre os termos?"
- Pois… claro… Mas como resolver?
  - E para vários idiomas (tal como os termos)
- Mantendo a filosofia semi-automática
  - Objectivo é ajudar o utilizador e não resolver o problema com 100% de garantia

Pesquisando Definições

- Para cada termo da base de dados procurar no corpus frases que "pareçam definições":
  - "TERMO é um XX que…"
  - "TERMO - XX que…"
  - "Entende-se por TERMO um XXX…"
  - Outras CENTENAS de padrões mais ou menos restritos
    - Compilados por Ana Sofia Pinto e Débora Oliveira (bolseiras)
- Apresentar os resultados ao utilizador
  - O utilizador valida os resultados e guarda na BD

Pesquisando Relações

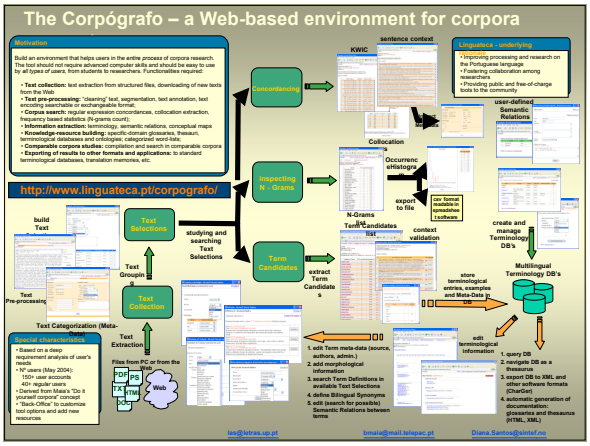
- Usando um método semelhante mas procurando padrões de 2 termos (A e B)
  - Hiponímia/Hiperonímia:
    - "…A é um B que…", "…A é um tipo de B"…
  - Meronímia/Holonímia:
    - "…os A de B…". "…A é uma parte de B que…"
  - "Causa / Efeito"
  - …
- Foram compilados manualmente várias dezenas de padrões indiciadores de relação para PT, EN, ES, FR…

A pesquisa de Relações…

- Para cada termo selecionado pelo utilizador:
  - Procurar frases em que o ocorra esse termo com outros termos que existam na BD
  - Para cada uma das frases seleccionadas verificar se obdecem aos padrões previamente compilados e que indiciam uma dada relação
  - Seleccionar as frases que obdecem aos padrões e mostrar ao utilizador para validação
  - Armazenar na BD as frase armazenadas pelo utilizador
- 100% Semi-automático…

Chegamos a Maio de 2004

- A primeira vez que o Corpógrafo é apresentado para um público numeroso:
  - LREC 2004
  - Contacto com Teresa Cabrè - UPF
- Até essa altura, ainda poucos utilizadores
  - Apenas na FLUP e em algumas Univ. BR
- Mas a grande conquista foi o inicio de uma pequena comunidade de interessados na FLUP em assuntos de Linguística Computacional



Ponto da situação nessa altura

- O sistema estava a ficar demasiado complexo:
  - Ainda muitas dependências a software externo
- Além disso:
  - As base de dados terminológicas estavam implementadas sobre sistema de ficheiros
    - O sistema estava a ficar demasiado lento
  - Cada vez mais interfaces de utilizador
    - difícil manter, de corrigir erros, de uniformizar aspecto
    - Mais de 60% do tempo de desenvolvimento
- Integração do Luís Cabral na equipa….

Tentando resolver alguns problemas…

- SAGI - projecto desenvolvido por Luis Cabral
  - Uniformização do aspecto das interfaces web
  - Possibilidade de gerir melhor processos
  - Ecrã de Espera para pesquisas mais lentas..
- Integração de um sistema de BD:
  - MySQL substituiu BD em ficheiros de texto
- Subsitituição do sistema CWB por MySQL
  - as pesquisas eram todas sobre texto não anotado
  - Permitia diminuir mais uma dependência externa

Preparação do Corpógrafo V2

- Essencialmente fazer as alterações técnicas mas não muitas a nível de funcionalidades para o utilizador
- Conseguiu-se:
  - um sistema mais estável e sustentável
  - um sistema mais agradável visualmente
  - um sistema mais rápido que permitiria mais utilizadores
- Sistema V2 lançado em finais de 2004
  - Grande explosão do Corpógrafo em número de utilizadores

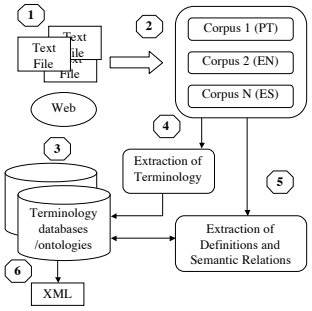
Preparação do Corpógrafo V3

- Os utilizadores queriam toda uma série de novas funcionalidades:
  - Pesquisas mais poderosas
  - Possibilidade de integrar objectos multimédia (vídeos, imagens, etc…) nas BD terminológicas
  - Formas mais simples de trabalhar
- Nós pretendiamos mais funcionalidades a nível da gestão de utilizadores. Simplicidade na
  - Criação de novas contas
  - Alteraração de parâmetros do Corpógrafo

Corpógrafo V3

- Versão lançada em meados de 2005
  - Pré-apresentada no CL 2005 (dois anos depois)
- Muitas pequenas funcionalidade novas
- Melhoria do algoritmo de pesquisa de terminologia
- Tentativa de lançar o código em licença GPL:
  - Pouco sucesso, infelizmente (já falarei sobre isto)
- Versão actual: 3.1 de Novembro de 2005
  - Algums "bugs" corrigidos (mas não todos)

Resumo do Corpógrafo V3

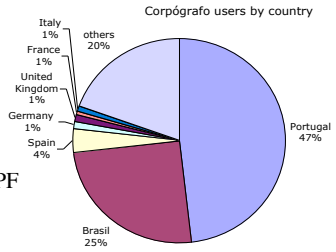


Quem são os utilizadores do Corpógrafo?

- Quem está a usar o Corpógrafo?
- O que é que estão a fazer com o Corpógrafo?
- Que segmentos?
  - Estudantes?
  - Tradutores?
  - Linguístas?
  - Engenheiros?
- Não sabemos muito bem:
  - gostava de poder ouvir alguns utilizadores que estejam aqui…

Número Globais

- Dados de Fevereiro de 2006
- 650 Inscrições
- Mais de 100 utilizadores "habituais"
- Número já subiram > 700
- Servidor na UPF



Por organização…

Users	Organization (country)
154	Univ. do Porto (PT)
49	Univ. de S. Paulo (BR)
39	Univ. do Minho (PT)
32	Univ. de Aveiro (PT)
31	Pontificia Univ. Cat. Rio Grande do Sul (BR)
14	Univ. Federal do Rio Grande do Sul (BR)
12	Univ. Nova de Lisboa (PT)
11	Univ. do Vale do Rio dos Sinos (BR)
11	Univ. de Lisboa (PT)
11	Univ. Salamanca (ES)
10	Univ. do Algarve (PT)
9	Linguateca (PT/NO)
267	Independent users, groups with 5 or less users

Projectos na FLUP

- FLUP: Faculdade de Letras da Univ. do Porto
  - Mestrado em Terminologia e Tradução (Prof. B. Maia)
- Projectos de Terminologia com
  - Engenharia Mecânica – Materiais Compósitos
  - Engenharia Electrotecnia – Telecomunicações móveis
  - Engenharia de Energias Alternativas
  - Geografia – Demografia da População
  - Geografia – Riscos Naturais
- Projectos feitos em parceria com outros grupos

Balanço dentro da FLUP

- Do ponto de vista pedagógico não posso afirmar nada muito porque não estou directamente envolvido
- Mas, do ponto de vista de um observador técnico e em 3 anos que acompanhei o projecto:
  - Criação de uma comunidade de utilizadores tecnicamente mais sensivel a assuntos de Linguística Computacional

Evolução na FLUP

- Para além das competências Linguística nativas, a comunidade adquiriu:
  - Sensibilidade para problemas técnicos
  - Sensibilidade para *escolha* e *utilização* de Corpora
  - Noção da importância da terminologia na SC
  - Compreensão dos novos papeis dos Linguístas / Tradutores na sociedade digital
  - Projectos interdisciplinares

Balanço do Corpógrafo

- Pontos Fortes:
  - Simples para o utilizador
  - Técnicas simples de PLN
  - Aplicações práticas de Corpora
  - Aprendizagem Colectiva
  - Criação de Comunidade
  - criação de recursos de terminologia
  - Apoio pedagógico
- Pontos Fracos:
  - Complexidade técnica do sistema
  - Dependência externas
  - Dificuldade em suportar todos os utilizadores
  - Dificuldade de instalação de novos servidores
  - Ainda pouca aplicação dos recursos produzidos

Aplicação dos recursos produzidos

- A grande prova pela qual o Corpografo ainda não passou:
  - Será que é possível utilizar os recursos produzidos pelo Corpógrafo para melhorar outras aplicações?
  - Será que os engenheiros poderão utilizar as terminologias / ontologias produzidas com o Corpógrafo em suas aplicações?
  - Que aplicações podem usufrui desses recursos?

Recursos em XML

```
<TU_ENTRY IID="38">
  <TU>axon</TU>
  <GEN_INFO lang="EN" iso_type="entrada terminológica" iso_adm="estandardizado" iso_reg="neutro"
     iso_freq="usado com frequência" iso_orig="empréstimo interdisciplinar"/>
...
  <MORF_INFO gender="U" number="U" animacy="U" pos="undef"/>
  ...
<DEF_INFO CORPUS="Neurons" FILE="undef">
     <DEFINITION>The axon is the main conducting unit of the neuron , capable of conveying electrical signals
     along distances that range from as short as 0.1 mm to as long as 2 m .</DEFINITION>
 </DEF_INFO>
...
 <TU_REL OTHER_IID="128" STRING="terminal button" REL="HOLO/MERO" ROLE="HOLO"
      OTHER_ROLE="MERO"/>
...
<TU_EQUIV OTHER_IID="2" STRING="axónio" TYPE="sinónimo"/>
<TU_EQUIV OTHER_IID="210" STRING="axónio" TYPE="sinónimo"/>
<TU_EQUIV OTHER_IID="337" STRING="assone" TYPE="sinónimo"/>
<TU_EQUIV OTHER_IID="468" STRING="axon" TYPE="sinónimo"/>
<TU_EQUIV OTHER_IID="496" STRING="axón" TYPE="sinónimo"/>
<TU_EQUIV OTHER_IID="639" STRING="axone" TYPE="sinónimo"/>
...
</TU_ENTRY>
```

Algumas ideias ainda não experimentadas

- Integrar os recursos num sistema de RI
  - Expansão multilingue de queries
  - Especialização / Generalização de queries
  - Sugestão de termos relacionados
- Usar os recursos para compilação de corpora
  - Comparável bilingue, Semelhante em domínio
- Utilização em visualização de informação
  - Topic maps, ontologias visuais
- Acopolamento a WordNet.PT
  - Criação de secções do WordNet mais especializadas

Corpógrafo V4?

- Ainda não há planos para o lançamento do V4
  - A equipa de desenvolvimento foi alterada
    - Luís Cabral: Porto -> Oslo
    - Luís Sarmento: Iniciei doutoramento
  - Projecto complexo com muitas dependências
    - e nem sempre totalmente documentado
  - Novo desenvolvedor ainda em adaptação:
    - António Silva
  - Prioridades actuais:
    - Código em licença GPL + documentação
    - Divisão de Carga por servidores

Conclusão

- Corpógrafo é um projecto multi-disciplinar
- Actualmente mais utilizado no ensino de terminologia, mas poderá ser utilizado para outros fins
- Utiliza técnicas simples de PLN
- Sistema complexo do ponto de vista informático: prioridade actual garantir sustentabilidade
- Tem permitido aos utilizadores e aos desenvolvedores aprender muito em conjunto

Obrigado

- Perguntas?
- Comentários?

Contactos:
Luís Sarmento - las@fe.up.pt
Belinda Maia - bmaia@mail.telepac.pt
Diana Santos - Diana.Santos@sintef.no